

# Proofreading of DNA Polymerase: a new kinetic model with higher-order terminal effects

Yong-Shun Song,<sup>1</sup> Yao-Gen Shu,<sup>2</sup> Xin Zhou,<sup>1</sup> Zhong-Can Ou-Yang,<sup>2</sup> and Ming Li<sup>1, a)</sup>

<sup>1)</sup>*School of Physical Science, University of Chinese Academy of Sciences*

<sup>2)</sup>*Institute of Theoretical Physics, Chinese Academy of Sciences*

(Dated: 10 May 2016)

The fidelity of DNA replication by DNA polymerase (DNAP) has long been an important issue in biology. While numerous experiments have revealed details of the molecular structure and working mechanism of DNAP which consists of both a polymerase site and an exonuclease (proofreading) site, there were quite few theoretical studies on the fidelity issue. The first model which explicitly considered both sites was proposed in 1970s' and the basic idea was widely accepted by later models. However, all these models did not systematically and rigorously investigate the dominant factor on DNAP fidelity, i.e, the higher-order terminal effects through which the polymerization pathway and the proofreading pathway coordinate to achieve high fidelity. In this paper, we propose a new and comprehensive kinetic model of DNAP based on some recent experimental observations, which includes previous models as special cases. We present a rigorous and unified treatment of the corresponding steady-state kinetic equations of any-order terminal effects, and derive analytical expressions for fidelity in terms of kinetic parameters under bio-relevant conditions. These expressions offer new insights on how the higher-order terminal effects contribute substantially to the fidelity in an order-by-order way, and also show that the polymerization-and-proofreading mechanism is dominated only by very few key parameters. We then apply these results to calculate the fidelity of some real DNAPs, which are in good agreements with previous intuitive estimates given by experimentalists.

PACS numbers: 87.10.Ed, 82.39.-k, 87.15.R-

Keywords: DNA polymerase; proofreading; kinetics; fidelity

---

<sup>a)</sup>Electronic mail: liming@ucas.ac.cn

## I. INTRODUCTION

Since the Watson-Crick base-pairing rules of double-strand DNA was established<sup>1</sup>, template-directed DNA replication became an important issue both in basic researches and application studies (e.g., Polymerase Chain Reaction ) in biology. The match between the incoming nucleotide dNTP and the template (i.e., the canonical Watson-Crick base pairing A-T and G-C) in the replication process plays a central role for any organism to maintain its genome stability, whereas mismatch (non-canonical base pairing like A-C) may introduce harmful genetic variations into the genome, and thus the error rate of replication must be kept very low. In living cells, the replication fidelity is controlled mainly by DNA polymerase (DNAP)<sup>2</sup> which catalyzes the template-directed DNA synthesis, and the fidelity of DNAP has been intensively studied since its discovery in 1950s'.<sup>3</sup>

Pioneering theoretical studies on this issue were done by J.Hopfield<sup>4</sup> and J.Ninio<sup>5</sup>. Regarding DNA replication approximately as a binary copolymerization process of matched nucleotides (denoted as A for convenience in the present paper) and mismatched nucleotides (denoted as B), they proposed independently the so-called kinetic proofreading mechanism which correctly points out that the replication fidelity is not determined thermodynamically by the free energy difference, but kinetically by the incorporation rate difference, between the match and the mismatch. This model, however, assumed that the proofreading occurs before nucleotide incorporation is accomplished (as illustrated in FIG. 1(a1)), which is not the case of real DNAPs. Structural and functional studies show that DNAP often has two parts. The basic part of all DNAPs is a synthetic domain (i.e., polymerase ) which binds the incoming dNTP and catalyzes its incorporation into the nascent ssDNA strand (called as primer below for convenience). Proofreading is performed by a second domain (i.e, exonuclease) which is not a necessary part of DNAP. This domain may much likely excise the just-incorporated mismatched nucleotide, once the mismatched terminus is transferred from the polymerase site into the exonuclease site by thermal fluctuation. The first model that explicitly invokes the exonuclease, referred to as Galas-Branscomb model (FIG. 1(b1)), was proposed by Galas et al.<sup>6</sup> and revisited by many other groups<sup>7-10</sup>. Many experimental studies gave consistent results to this model<sup>11-13</sup>. Recently, improved experimental techniques revealed more details of the synthesizing and proofreading processes<sup>14,15</sup>, and several detailed kinetic models have been proposed<sup>15-17</sup>. However, all these models are based on the original

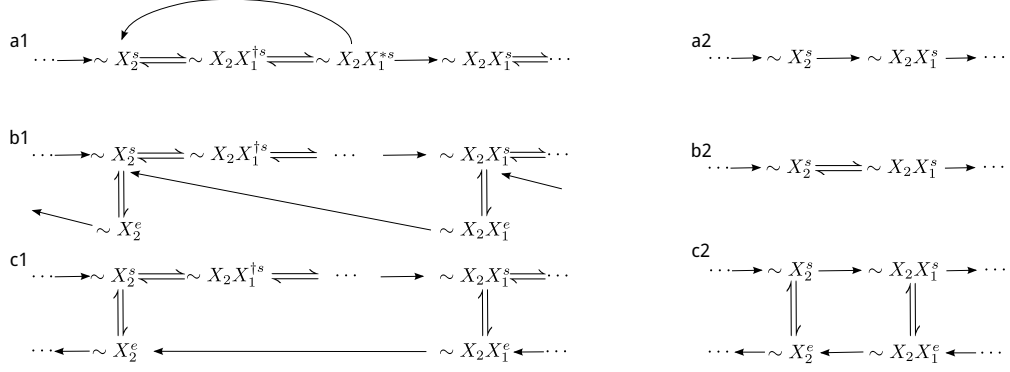


FIG. 1. Three different proofreading mechanisms of DNAP. For simplification, matched or mismatched dNTP is represented by  $A$  or  $B$  respectively throughout this paper.  $X_i (i = \dots, 1, 2, \dots)$  denotes either  $A$  or  $B$ . The superscript  $s$  or  $e$  means that the primer terminus is in the polymerase (i.e., synthetic) site or the exonuclease site, respectively.  $\sim X_2 X_1^{\dagger s}$  denotes the state dNTP binding to DNAP before DNAP undergoes further conformation change. (a1) The original kinetic proofreading mechanism.  $\sim X_2 X_1^{*s}$  represents one or more high-energy intermediate states which dissociate much faster for  $B$  than for  $A$ . This proofreading occurs before the nucleotide is covalently incorporated into the primer. (b1) Brief sketch of the Galas-Branscomb model. (c1) An alternative exonuclease proofreading mechanism proposed in this paper. Considering only the exact calculation of the fidelity, one can simplify these schemes under steady-state conditions, i.e., (a1) can be simplified as minimal scheme (a2), (b1) simplified as (b2), and (c1) simplified as (c2).<sup>1</sup>

simple Galas-Branscomb model and many important details such as higher-order neighbor effects of the primer terminus are not considered systematically<sup>17</sup> (see later sections). In particular, recent experimental works on phi29 DNA polymerase<sup>18,19</sup> revealed more details about the working mechanism of DNAP, highlighting the importance of the forward and backward translocation steps which were absent from the Galas-Branscomb models. Considering this point, as well as many other structural<sup>20–24</sup> and kinetic<sup>8,12,18,19,25</sup> experimental results, we propose a comprehensive reaction scheme of DNAPs as shown in FIG. 2.

There are several key features of this scheme. First, the template-primer duplex binds to DNAP and forms two types of complexes. In the ‘polymerase type’, the 3’ terminus of the

<sup>1</sup> In Section III, the replication fidelity is defined as the ratio between the steady-state flux  $J_A$  and  $J_B$ . In order to calculate such steady-state flux-flux ratios, one can map the original schemes to much simplified versions. For instance, any multistep pathway without branches can be mapped to a single-step pathway. Thus one obtains the much simplified schemes (a2), (b2) and (c2). On the other hand, in many kinetic assays of the DNAP reactions, multiple steps in the same pathway cannot be identified individually. In such cases, minimal schemes like (a2), (b2) and (c2) are directly used to analyze the experimental data.

primer is located at the polymerase site. In the ‘exonuclease type’, the primer terminus is unzipped from the duplex and transferred to the exonuclease site. For the ‘polymerase type’ complexes, two substates were experimentally observed<sup>18,19</sup>. One is the pre-translocation state of DNAP in which the dNTP binding site is occupied by the primer terminus. The other is the post-translocation state in which the DNAP translocates forward (relative to the template) to expose the binding site to the next dNTP. DNAP can rapidly switch between these two states. Correspondingly, one can assume two substates of DNAP in the ‘exonuclease type’ complexes, though there are not sufficient experimental evidences. One is the pre-translocation state in which the exonuclease site is occupied when the primer terminus is transferred from the polymerase site. The other is the post-translocation state in which the exonuclease site is exposed after the nucleotide excision while the newly-formed primer terminus does not return to the polymerase site.

Second, once the incoming dNTP is incorporated into the primer, the DNAP can either translocate forward to the post-translocation state and bind a new dNTP in the polymerase site, or it pauses at the pre-translocation state and the primer terminus is unzipped from the duplex and transferred to the exonuclease site (the terminus can switch between the two sites without being excised<sup>19</sup>). The large distance about  $30 - 40 \text{ \AA}$ <sup>20–24</sup> between the two sites implies that more than one nucleotides of the primer terminus must be unzipped, and thus the stability of the entire terminal region may put an impact on the unzipping probability of the primer terminus. Such neighbor effects, as well as other types of neighbor effects, can be very significant for the replication fidelity and should be taken account of in the kinetic models (details see later sections).

Third, the exonuclease site can only excise the terminal nucleotide. What happens after the cleavage is not clear yet<sup>27</sup>. Here we propose two possible pathways, which are denoted as Model I and Model II in FIG. 2. In Model I, DNAP undergoes a backward translocation and the primer terminus can either be excised processively, or be transferred back to the polymerase site (at the pre-translocation state). In Model II, the primer terminus is directly transferred back to the polymerase site (at the post-translocation state). FIG. 2 can be further simplified as FIG. 3, considering that the addition of dNTP in the polymerase site is almost irreversible (i.e., the product PPi of the polymerization reaction is often released irreversibly under physiological conditions).

One can also reasonably assume that the translocation of DNAP in ‘polymerase type’

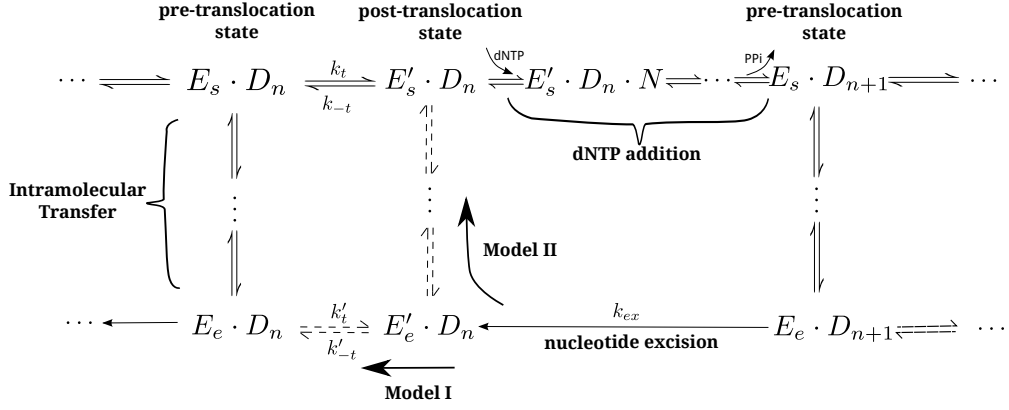


FIG. 2. The reaction scheme for DNAP.  $E$ : DNA polymerase;  $D_n$ : the state of the primer,  $n$  being the length of the primer;  $N$ : dNTP.  $E_s \cdot D_n$  and  $E'_s \cdot D_n$ : the ‘polymerase type’ complex when DNAP is in the pre-translocation state and post-translocation state, respectively.  $E_e \cdot D_n$  and  $E'_e \cdot D_n$ : the ‘exonuclease type’ complex when DNAP is in the pre-translocation state and post-translocation state, respectively. A free dNTP can bind to DNAP when the complex is at the post-translocation state  $E'_s \cdot D_n$ . When the dNTP is incorporated into the primer, the complex will return to the pre-translocation state  $E_s \cdot D_{n+1}$ . The primer terminus may be unzipped from the duplex and transferred to the exonuclease site. Model I and Model II indicate two possible pathways after the nucleotide excision in the exonuclease site.

complex is in a rapid equilibrium. In biochemical experimental studies such as steady-state kinetic assays<sup>15,25</sup>, the translocation cannot be observed (for comparison, the subsequent dNTP binding can be clearly observed). In other words, the two substates can not be identified individually, indicating there exists a rapid equilibrium between them. Thus one does not need to distinguish between the pre-translocation and the post-translocation states. Under such an approximation, Model II can be reduced to the Galas-Branscomb model as shown in FIG. 1(b1), while Model I is reduced to FIG. 1(c1).

Although Model II were widely accepted, there is no direct experimental evidence to exclude Model I. Moreover, it has been found that the ssDNA binding to the exonuclease site can be processively excised<sup>25</sup>, indicating that more than one nucleotide bind to the exonuclease site (e.g., three nucleotides bind to the exonuclease site for Polymerase I KF<sup>28</sup>) and removing the terminal nucleotide may trigger backward translocation of DNAP for the subsequent excisions. So we will discuss both models in this paper, but put a focus on Model I due to the following technique consideration. Kinetic proofreading models like

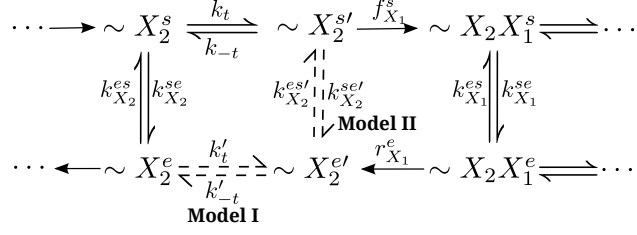


FIG. 3. The simplified reaction scheme.  $X^s$ ,  $X^e$  (or  $X^{s'}$ ,  $X^{e'}$ ): pre-translocation (or post-translocation) state of DNAP when the primer terminus is in the synthetic(s) site or the exonuclease(e) site respectively. When the primer terminus is in the exonuclease site, one does not need to distinguish between  $\sim A^e(\sim B^e)$ . However, it's still convenient to use  $\sim A^e(\sim B^e)$  to denote the immediate state when the terminus switches back to the polymerase site. By setting all the excision rates equal to  $r^e$ , we obtain the models for real DNAPs. Under the steady-state conditions, the effective rate of dNTP addition can be expressed as  $f_{X_1}^s = k_p[X_1]$ , where  $k_p$  is an effective quasi-first-order rate constant,  $[X_1]$  is the concentration of the incoming dNTP (to calculate the intrinsic fidelity, one often sets  $[A] = [B]$ ). All other kinetic parameters in this figure are effective parameters which are combinations of the original rate constants in FIG. 2.

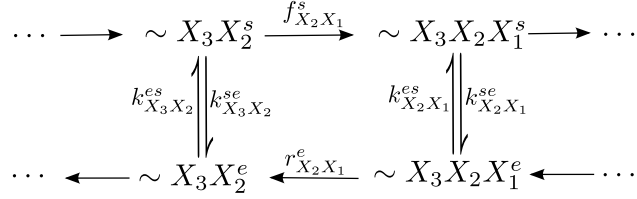


FIG. 4. The minimal scheme of the first-order proofreading model.

FIG. 1(a1) or (a2) are irreversible reactions, so the corresponding kinetic equations are always closed (i.e., of finite number) and can be rigorously calculated. The Galas-Branscomb models like FIG. 1(b1) or (b2), however, are seemingly reversible, and the corresponding kinetic equations are always unclosed and hierarchically coupled, which is hard to solve. Fortunately, a general rigorous treatment for such problems has been established recently by us<sup>29</sup>, and this method can be directly applied to Model II (some important results are given in Appendix C). For Model I like FIG. 1(c1) or (c2), however, the above methods are inapplicable and new method should be developed, which will be a focus of later sections.

This paper is organized as follows. Section II introduces the basic theory of the steady-state kinetics of Model I (the minimal scheme FIG. 1(c2)) including higher-order neighbor

effects. In Section III, we discuss the replication fidelity problem of DNAP with either Model I or Model II (FIG. 1(c2),(b2) respectively). While it's hard to analytically calculate the fidelity in terms of the kinetic parameters from the basic kinetic theory, we introduce an alternative method (infinite-state Markov chain) for the calculation and show numerically its equivalence to our basic theory. With this method, analytical expressions for fidelity are obtained under the so-called biologically-relevant conditions. We further show that Model I and II give exactly the same expressions which offer an intuitive understanding of the higher-order neighbor effects on the fidelity. In Section IV, we will apply these results to discuss the fidelity problem of some real DNAPs.

## II. BASIC KINETIC THEORY OF PROOFREADING MODEL I

It has been shown that the terminal mismatch and even the penultimate mismatch at the primer terminus will greatly reduce the addition rate of the next dNTP, compared with the case that a match is at the same position<sup>8,30,31</sup>. This means that some rate constants in FIG. 2 depends on the states ( $A$  or  $B$ ) of the few consecutive base pairs at the terminal region, i.e., there does exist higher-order neighbor effects (referred to as terminal effects in this paper) in DNA replication. Thus the zero-order terminal model shown in FIG. 3 is not appropriate and higher-order models like FIG. 4 or FIG. 6 are required. Below we demonstrate how to rigorously treat the steady-state kinetics of such models. To proceed, we note first that each step in the reaction scheme may have terminal effect but of different order. For instance, the addition rate may be of first order while the transfer rate may be of zero order, which is a special case of the general first-order scheme FIG. 4 (by putting  $k_{AX_1}^{se} = k_{BX_1}^{se}$ ). Similarly, reaction schemes with kinetic parameters up to  $sth$  order can be included in the general  $sth$ -order scheme.

### A. First-order proofreading model

In this subsection, we will discuss the general first-order proofreading model FIG. 4 to demonstrate the basic ideas of our approach. Following the same logic of Ref. 29, we use  $P_{X_n \cdots X_1}^s$  to denote the occurrence probability of the terminal sequence  $X_n \cdots X_1$  in the synthetic (polymerase) site,  $P_{X_n \cdots X_1}^e$  to denote the occurrence probability of  $X_n \cdots X_1$  in

the exonuclease site,  $X_i = A, B$ .  $N_{X_n \dots X_2 X_1}$  is defined as the total number of sequence  $X_n \dots X_2 X_1$  appearing in the primer chain.

The overall incorporation rate of sequence  $X_n \dots X_2 X_1$  ( $n \geq 2$ ) is defined as,

$$\dot{N}_{X_n \dots X_2 X_1} \equiv J_{X_n \dots X_2 X_1} = J_{X_n \dots X_2 X_1}^s + J_{X_n \dots X_2 X_1}^e, \quad (1)$$

where  $J_{X_n \dots X_2 X_1}^s = f_{X_2 X_1}^s P_{X_n \dots X_2}^s$ ,  $J_{X_n \dots X_2 X_1}^e = -r_{X_2 X_1}^e P_{X_n \dots X_2 X_1}^e$ .

The kinetic equations of  $P_{X_n \dots X_2 X_1}^m$  ( $n \geq 1, m = s, e$ ) can be written as,

$$\dot{P}_{X_n \dots X_2 X_1}^s = J_{X_n \dots X_2 X_1}^s - \tilde{J}_{X_n \dots X_2 X_1}^s - J_{X_n \dots X_2 X_1}^{se}, \quad (2a)$$

$$\dot{P}_{X_n \dots X_2 X_1}^e = J_{X_n \dots X_2 X_1}^e - \tilde{J}_{X_n \dots X_2 X_1}^e + J_{X_n \dots X_2 X_1}^{se}, \quad (2b)$$

where,  $\tilde{J}_{X_n \dots X_1}^s = J_{X_n \dots X_1 A}^s + J_{X_n \dots X_1 B}^s$ ,  $\tilde{J}_{X_n \dots X_1}^e = J_{X_n \dots X_1 A}^e + J_{X_n \dots X_1 B}^e$ ,  $J_{X_n \dots X_2 X_1}^{se} = k_{X_2 X_1}^{se} P_{X_n \dots X_2 X_1}^s - k_{X_2 X_1}^{es} P_{X_n \dots X_2 X_1}^e$ . We also have  $P_{X_i \dots X_1}^s = P_{A X_i \dots X_1}^s + P_{B X_i \dots X_1}^s$ ,  $J_{X_i \dots X_1}^s = J_{A X_i \dots X_1}^s + J_{B X_i \dots X_1}^s$  ( $i \geq 1$ ) and so on.

For example,

$$\begin{aligned} \dot{P}_{AB}^s &= f_{AB}^s (P_{AA}^s + P_{BA}^s) - (f_{BA}^s + f_{BB}^s) P_{AB}^s - k_{AB}^{se} P_{AB}^s + k_{AB}^{es} P_{AB}^e \\ &= J_{AB}^s - (J_{ABA}^s + J_{ABB}^s) - J_{AB}^{se}, \end{aligned} \quad (3a)$$

$$\begin{aligned} \dot{P}_{AB}^e &= -r_{AB}^e P_{AB}^e + r_{BA}^e P_{ABA}^e + r_{BB}^e P_{ABB}^e + k_{AB}^{se} P_{AB}^s - k_{AB}^{es} P_{AB}^e \\ &= J_{AB}^e - (J_{ABA}^e + J_{ABB}^e) + J_{AB}^{se}. \end{aligned} \quad (3b)$$

The steady state is defined as  $\dot{P}_{X_n \dots X_2 X_1}^s = 0$  and  $\dot{P}_{X_n \dots X_2 X_1}^e = 0$  for any  $n \geq 1$ . To rigorously solve these coupled equations, we extend the logic of Ref. 29 and propose the following factorization conjecture:

$$P_{X_n \dots X_2 X_1}^m = \prod_{i=3}^n P_{X_i X_{i-1}}^s \left[ \prod_{i=3}^n P_{X_{i-1}}^s \right]^{-1} P_{X_2 X_1}^m, \quad n \geq 3, \quad m = s, e. \quad (4)$$

For example,  $P_{X_3 X_2 X_1}^s = P_{X_3 X_2}^s P_{X_2 X_1}^s / P_{X_2}^s$ ,  $P_{X_3 X_2 X_1}^e = P_{X_3 X_2}^e P_{X_2 X_1}^e / P_{X_2}^e$  (correspondingly, one also has  $J_{X_3 X_2 X_1}^s = P_{X_3 X_2}^s J_{X_2 X_1}^s / P_{X_2}^s$  and  $J_{X_3 X_2 X_1}^e = P_{X_3 X_2}^e J_{X_2 X_1}^e / P_{X_2}^e$ ). The validity of these factorization conjectures can be numerically tested by Monte Carlo simulation (using the Gillespie algorithm<sup>32</sup>), as shown in FIG. 5.

By this factorization conjecture, the original unclosed equations can be reduced to the



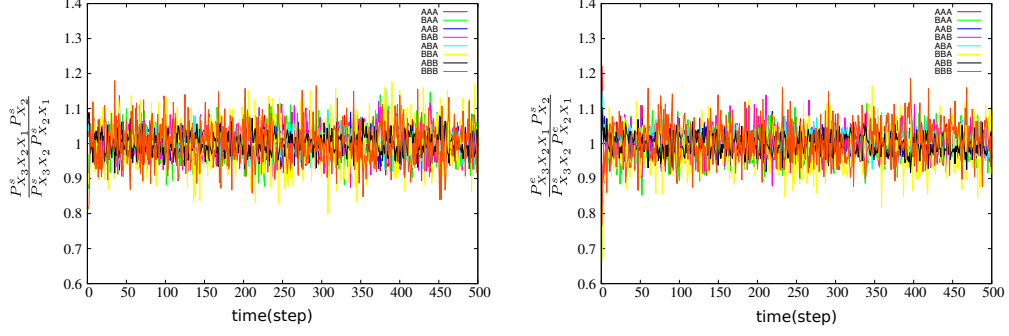


FIG. 5. Simulation verification of the factorization conjecture of the first-order proofreading model, with illustrative rate parameters (in unit  $s^{-1}$ )  $f_{AA}^s = 8$ ,  $f_{AB}^s = 6$ ,  $f_{BA}^s = 4$ ,  $f_{BB}^s = 2$ ,  $r_{AA}^e = 1$ ,  $r_{AB}^e = 1$ ,  $r_{BA}^e = 2$ ,  $r_{BB}^e = 1$ ,  $k_{AA}^{se} = 1$ ,  $k_{AB}^{se} = 6$ ,  $k_{BA}^{se} = 1$ ,  $k_{BB}^{se} = 6$ ,  $k_{AA}^{es} = 1$ ,  $k_{AB}^{es} = 3$ ,  $k_{BA}^{es} = 1$ ,  $k_{BB}^{es} = 4$ . Averaged over 10,000 samples.

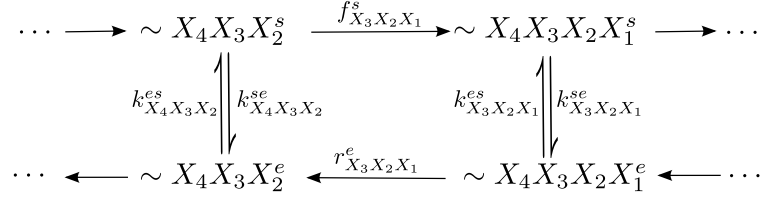


FIG. 6. The second-order proofreading reaction scheme.

following closed equations of the eight basic variables  $P_{X_2 X_1}^m (m = s, e)$ :

$$\begin{aligned}
 J_{BA}^e - J_{AB}^e &= J_B^{se}, & J_{BA}^s - J_{AB}^s &= J_A^{se}, \\
 \frac{J_{AA}^s - J_{AA}^{se}}{J_{BA}^s - J_{BA}^{se}} &= \frac{P_{AA}^s}{P_{BA}^s}, & \frac{J_{AB}^s - J_{AB}^{se}}{J_{BB}^s - J_{BB}^{se}} &= \frac{P_{AB}^s}{P_{BB}^s}, \\
 \frac{J_{AA}^e + J_{AA}^{se}}{J_{BA}^e + J_{BA}^{se}} &= \frac{P_{AA}^e}{P_{BA}^e}, & \frac{J_{AB}^e + J_{AB}^{se}}{J_{BB}^e + J_{BB}^{se}} &= \frac{P_{AB}^e}{P_{BB}^e}, \\
 J_A^{se} + J_B^{se} &= 0, & \sum_{X,Y=A,B} (P_{XY}^s + P_{XY}^e) &= 1.
 \end{aligned} \tag{5}$$

## B. Second-order proofreading model

Second-order terminal effects have been observed for some DNAPs where the penultimate mismatch at the terminus can affect the next nucleotide incorporation<sup>26,31</sup>. In this section, we extend the method of the preceding subsection to the second-order model shown in FIG. 6.

Similar to the first-order model, we have,

$$\dot{N}_{X_n \dots X_2 X_1} \equiv J_{X_n \dots X_2 X_1} = J_{X_n \dots X_2 X_1}^s + J_{X_n \dots X_2 X_1}^e \quad (n \geq 3), \quad (6)$$

where  $J_{X_n \dots X_3 X_2 X_1}^s = f_{X_3 X_2 X_1}^s P_{X_n \dots X_3 X_2}^s$ ,  $J_{X_n \dots X_3 X_2 X_1}^e = -r_{X_3 X_2 X_1}^e P_{X_n \dots X_3 X_2 X_1}^e$ .

The kinetic equations for  $P_{X_n \dots X_3 X_2 X_1}^m$  ( $n \geq 1, m = s, e$ ) can be written as,

$$\dot{P}_{X_n \dots X_3 X_2 X_1}^s = J_{X_n \dots X_3 X_2 X_1}^s - \tilde{J}_{X_n \dots X_3 X_2 X_1}^s - J_{X_n \dots X_3 X_2 X_1}^{se}, \quad (7a)$$

$$\dot{P}_{X_n \dots X_3 X_2 X_1}^e = J_{X_n \dots X_3 X_2 X_1}^e - \tilde{J}_{X_n \dots X_3 X_2 X_1}^e + J_{X_n \dots X_3 X_2 X_1}^{se}, \quad (7b)$$

where  $J_{X_n \dots X_3 X_2 X_1}^{se} = k_{X_3 X_2 X_1}^{se} P_{X_n \dots X_3 X_2 X_1}^s - k_{X_3 X_2 X_1}^{es} P_{X_n \dots X_3 X_2 X_1}^e$ .

Under steady-state conditions  $\dot{P}_{X_n \dots X_3 X_2 X_1}^s = \dot{P}_{X_n \dots X_3 X_2 X_1}^e = 0$ , we proposed the following factorization conjecture:

$$P_{X_n \dots X_3 X_2 X_1}^m = \prod_{i=4}^n P_{X_i X_{i-1} X_{i-2}}^s \left[ \prod_{i=4}^n P_{X_{i-1} X_{i-2}}^s \right]^{-1} P_{X_3 X_2 X_1}^m, \quad n \geq 4, \quad m = s, e, \quad (8)$$

which can be tested by Monte Carlo simulations (results not shown here).

Therefore, we obtain the following closed equations for the second-order proofreading model:

$$\begin{aligned} J_{X\bar{X}}^s - \tilde{J}_{X\bar{X}}^s &= J_{X\bar{X}}^{se}, \quad J_{\bar{X}XX}^s - J_{XX\bar{X}}^s = J_{XX}^{se}, \\ \frac{J_{AXY}^s - J_{AXY}^{se}}{J_{BXY}^s - J_{BXY}^{se}} &= \frac{P_{AXY}^s}{P_{BXY}^s}, \quad J_{XX\bar{X}} = J_{\bar{X}XX}, \\ \frac{J_{AXY}^e + J_{AXY}^{se}}{J_{BXY}^e + J_{BXY}^{se}} &= \frac{P_{AXY}^s}{P_{BXY}^s}, \quad J_{AB} = J_{BA}, \\ \sum (P_{XYZ}^s + P_{XYZ}^e) &= 1, \quad X, Y, Z = A, B, \end{aligned} \quad (9)$$

where  $\bar{X}$  differs from  $X$ .

Some experiments<sup>8,33</sup> show that up to 4 base pairs at the primer terminus may have apparent effects on the incorporation rates of the next nucleotide. For such cases, one should generalize the above method to include these higher-order terminal effects. The generalization to sth-order model is straightforward and details are not given here.

### III. THE FIDELITY PROBLEM OF DNA REPLICATION BY DNAP

In this section, we discuss the fidelity problem of DNAP. In principle, one can define the fidelity naturally as the ratio of matches over mismatches incorporated into the primer. However, it's difficult to directly measure this fidelity in experiments and some indirect methods

were developed. One of the common used methods is the forward mutation assay<sup>13,34–36</sup> which scores the replication errors indirectly by counting the phenotype change rate of the bacterial hosts transfected by reporter gene DNA. Other frequently used methods are steady-state<sup>37–39</sup> or pre-steady state<sup>12,40–42</sup> kinetic assays which investigate the kinetics of DNA replication and calculate the replication fidelity indirectly based on the theoretical models. The basic ideas of these two approaches differ, but the obtained fidelity are often of similar order of magnitudes. For example, the average fidelity of *Sulfolobus solfataricus* P2 DNA polymerase IV (Dpo4) is about  $1.3 \times 10^2$  to  $3.3 \times 10^3$  using steady-state kinetic method<sup>43</sup>, which agrees with  $1.5 \times 10^2$  given by forward mutation assay<sup>36</sup>. In general, for most proofreading-proficient DNAPs, the fidelity *in vitro* is about  $10^6 - 10^7$  with a contribution by exonuclease proofreading of  $10^1 - 10^2$ .<sup>3,44</sup>

In this paper, we only discuss the kinetic-based fidelity, since it can be rigorously defined and calculated within the framework of our basic theory. Here we define the fidelity as  $\phi = N_A/N_B$ .  $N_A$  is the total number of incorporated matches in the primer,  $N_B$  is the total number of mismatches. Once the steady-state kinetic equations such as Eqs. (5) or Eqs. (9) are solved numerically or analytically, the total flux  $J_A (= J_A^s + J_A^e)$ ,  $J_B (= J_B^s + J_B^e)$  can be calculated. Since  $\dot{N}_A = J_A$ ,  $\dot{N}_B = J_B$ , and  $d(N_A/N_B)/dt = 0$  (in steady state), we can calculate the replication fidelity exactly by  $\phi = N_A/N_B = J_A/J_B$ . In particular, the analytical solutions to the kinetic equations are quite useful for further experimental and theoretical studies. However, it's often impossible to solve the kinetic equations analytically. To circumvent this problem, we introduce below an alternative method, the infinite-state Markov chain method<sup>45</sup>, to calculate  $\phi$ . This method has already been used for higher-order copolymerization by us (see the supplementary of Ref. 29) and can be readily extended to the exonuclease proofreading schemes.

### A. The infinite-state Markov chain method for exonuclease proofreading

To calculate the fidelity, we begin with the first-order proofreading scheme which can be rewritten as a branching model shown in FIG. 7.

The steady-state growth of primer can be completely characterized by four groups of

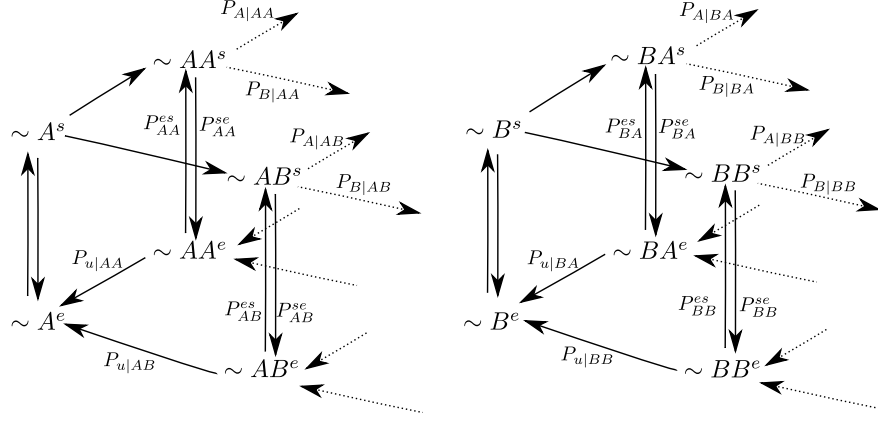


FIG. 7. Branching model for the first-order polymerization and excision.

transition probabilities:

$$P_{X|X_2X_1} \equiv f_{X_1X}^s / (f_{X_1X}^s + f_{X_1\bar{X}}^s + k_{X_2X_1}^{se}), \quad P_{X_2X_1}^{se} \equiv k_{X_2X_1}^{se} / (f_{X_1X}^s + f_{X_1\bar{X}}^s + k_{X_2X_1}^{se}),$$

$$P_{X_2X_1}^{es} \equiv k_{X_2X_1}^{es} / (r_{X_2X_1}^e + k_{X_2X_1}^{es}), \quad P_{u|X_2X_1} \equiv 1 - P_{X_2X_1}^{es} = r_{X_2X_1}^e / (r_{X_2X_1}^e + k_{X_2X_1}^{es}).$$

We also employ the idea of ‘cycle completion’<sup>45</sup>, since any incorporated nucleotide (either A or B) has a chance to be excised, only those not being excised account for the final composition of the primer. Thus the fidelity for the first-order terminal model can be defined as,

$$\phi \equiv \frac{Q_{AA} + Q_{BA}}{Q_{AB} + Q_{BB}}, \quad (10)$$

where  $Q_{X_2X_1}$  is the probability that  $X_1$  is added to the terminal  $X_2$  and never being excised, satisfying  $Q_{AA} + Q_{AB} + Q_{BA} + Q_{BB} = 1$ .  $Q_{X_2X_1}$  can be explicitly expressed as  $Q_{X_2X_1} \equiv \hat{P}_{X_2X_1} P_{nuX_2X_1}$ , where  $\hat{P}_{X_2X_1}$  is the probability that adding  $X_1$  to the terminal  $X_2$ ,  $P_{nuX_2X_1}$  is the probability of the terminal  $X_2X_1$  never being excised. The absolute values of  $\hat{P}_{X_2X_1}$  are not known *a priori*, but the following equalities obviously hold:

$$\frac{\hat{P}_{AA}}{\hat{P}_{AB}} = \frac{P_{A|AA}}{P_{B|AA}} = \frac{P_{A|BA}}{P_{B|BA}} = \frac{f_{AA}^s}{f_{AB}^s}, \quad \frac{\hat{P}_{BA}}{\hat{P}_{BB}} = \frac{P_{A|AB}}{P_{B|AB}} = \frac{P_{A|BB}}{P_{B|BB}} = \frac{f_{BA}^s}{f_{BB}^s}. \quad (11)$$

Considering the fact that the number of AB should equal to the number of BA in the copolymer chain, we have the following intrinsic constraint:

$$Q_{AB}(= \hat{P}_{AB} P_{nuAB}) = Q_{BA}(= \hat{P}_{BA} P_{nuBA}). \quad (12)$$

To calculate  $P_{nuX_2X_1}$ , we define  $P_{euX_2X_1} \equiv 1 - P_{nuX_2X_1}$  as the probability of the terminal  $X_2X_1$  ever being excised.  $P_{euX_2X_1}$  satisfy the following iterative equations (details can be

found in Appendix A):

$$P_{euX_2X_1} = \frac{\hat{P}_{u|X_2X_1}}{P_{X_2X_1}^{se} P_{X_2X_1}^{es}} \left( \frac{1}{1 - (\hat{P}_{A|X_2X_1} P_{euX_1A} + \hat{P}_{B|X_2X_1} P_{euX_1B}) P_{X_2X_1}^{es}} - \frac{1}{T_{X_2X_1}} \right). \quad (13)$$

Here,  $T_{X_2X_1} = 1/(1 - P_{X_2X_1}^{se} P_{X_2X_1}^{es})$ ,  $\hat{P}_{u|X_2X_1} = P_{u|X_2X_1} P_{X_2X_1}^{se} T_{X_2X_1} = \hat{r}_{X_2X_1}^s / (f_{X_1A}^s + f_{X_1B}^s + \hat{r}_{X_2X_1}^s)$ ,  $\hat{P}_{X|X_2X_1} = P_{X|X_2X_1} T_{X_2X_1} = f_{X_1X}^s / (f_{X_1X}^s + f_{X_1\bar{X}}^s + \hat{r}_{X_2X_1}^s)$ ,  $\hat{r}_{X_2X_1}^s = k_{X_2X_1}^{se} r_{X_2X_1}^e / (r_{X_2X_1}^e + k_{X_2X_1}^{es})$ .

Once  $P_{euX_2X_1}$  are solved,  $\hat{P}_{X_2X_1}$  can then be calculated by combining Eq. (11) and Eq. (12), and the fidelity  $\phi$  can be obtained by Eq. (10). Numerical calculations show that  $\phi$  obtained in this approach is identical to that given by the steady-state kinetic equations Eqs. (5), which provides a verification of our kinetic approach. The same logic can be extended to any higher-order models.

## B. Approximation of $\phi$ under bio-relevant conditions

In TABLE I, we list experimental values of some kinetic parameters for some real DNAPs. There exists huge difference in the order of magnitudes of the parameters of the same DNAP. For example, addition of matched nucleotide at the polymerase site is very fast, and always much faster than mismatch addition. This enables us to suggest reasonable approximations (so-called bio-relevant conditions in this paper) to simplify the above calculation and obtain explicit mathematical expressions of  $\phi$  in terms of some key parameters. For any higher-order models (say,  $h^{th}$ -order model), we propose

$$(a) \underbrace{f_{AAA\dots}^s}_{h+1} \gg \underbrace{f_{AAA\dots B}^s}_h, \text{ which leads to } P_{A|X \underbrace{AAA\dots}_h} \gg P_{B|X \underbrace{AAA\dots}_h}.$$

This means that the overall nucleotide incorporation is dominated by the addition of  $A$  and the occurrence probability of  $B$  in the primer is negligible. This highly efficient discrimination between  $A$  and  $B$  is executed by the polymerase site.

$$(b) \underbrace{f_{AAA\dots}^s}_{h+1} \gg k_X^{se} \underbrace{f_{AAA\dots}^s}_h (> \hat{r}_X^s \underbrace{f_{AAA\dots}^s}_h), \text{ which leads to } P_{A|X \underbrace{AAA\dots}_h} \gg P_X^{se} \underbrace{f_{AAA\dots}^s}_h (> \hat{P}_{u|X \underbrace{AAA\dots}_h}).$$

This can be achieved at appropriate concentration of dNTP (notice that  $\underbrace{f_{AAA\dots}^s}_{h+1}$  is proportional to dNTP concentration). It means that the matched terminus can be rapidly extended by the next match, instead of being transferred to the exonuclease site and excised. This ensures that the primer growth is dominated by match extension in the polymerase site and

TABLE I. Experimental values of kinetic parameters of some real DNAPs

rate parameter		T7 <sup>25,30</sup> ( $s^{-1}$ )	pol $\gamma$ <sup>26,31</sup> ( $s^{-1}$ )	Pol III <sup>46,47</sup> ( $s^{-1}$ )	T4 <sup>48</sup> ( $s^{-1}$ )	phi29 DNAP <sup>19,49,50</sup> ( $s^{-1}$ )
first order	$f_{AA}^s$	250 <sup>a</sup>	3900-5700 <sup>c</sup>	370 <sup>e</sup>	600	680 <sup>g</sup>
	$f_{AB}^s$	0.002 <sup>a</sup>	0.023-1.6 <sup>c</sup>	$(0.16-2.1) \times 10^{-3e}$	*	$10^{-4}-10^{-6g}$
	$k_{AA}^{se}$	0.2 <sup>b</sup>	$>\sim 0.05^d$	0.015 <sup>f</sup>	1	11.54 <sup>i</sup>
	$k_{AB}^{se}$	2.3 <sup>b</sup>	$>\sim 0.4^d$	0.038 <sup>f</sup>	5	—
	$k_{AA}^{es}$	714 <sup>b</sup>	$>\sim 39^d$	—	20	10.48 <sup>i</sup>
	$k_{AB}^{es}$	714 <sup>b</sup>	—	—	—	—
	$r^e$	896 <sup>b</sup>	$>\sim 39^d$	280 <sup>f</sup>	100	500 <sup>h</sup>
second order	$k_{ABA}^{se}$	—	$>\sim 3^d$	—	—	—
	$k_{ABA}^{es}$	—	—	—	—	—
	$f_{ABA}^s$	0.012 <sup>a</sup>	0.1 <sup>c</sup>	—	—	$10^{-3} - 10^{-4g}$
	$f_{BAA}^s$	—	2.7 <sup>c</sup>	—	—	—

Polymerization parameters are all scaled to the standard dNTP concentration  $100\mu M$ .

— means the data were not found. \* means the data is too small to measure.

<sup>a</sup> from Table II of Ref. 30.

<sup>b</sup> from Ref. 25.

<sup>c</sup> from Ref. 31, dNTP concentration is set as  $100\mu M$  for holoenzyme. The listed values differ for different base pairs (matched or mismatched). This type of sequence effect is beyond the scope of this paper and not discussed here.

<sup>d</sup> estimated from the combined kinetic parameters from Ref. 26.

<sup>e</sup> values for the holoenzyme, from Ref. 46.

<sup>f</sup> estimated by pre-steady state measurements of purified  $\epsilon$  subunit, from Ref. 47.

<sup>g</sup> from Table I and Table II of Ref. 49 in  $Mg^{2+}$ -activated polymerization.

<sup>h</sup> from Ref. 50.

<sup>i</sup> from Ref. 19.

the introduction of exonuclease proofreading pathway nearly does not change the overall growth velocity.

(c)  $\hat{r}_{\underbrace{AAA\dots}_{h-i+1} \underbrace{BAAA\dots}_{i-1}}^s \gtrsim f_{\underbrace{AAA\dots}_{h-i} \underbrace{BAAA\dots}_i}^s$  for  $0 < i \leq m$  (which leads to  $\hat{P}_u|_{\underbrace{AAA\dots}_{h-i+1} \underbrace{BAAA\dots}_{i-1}} (\equiv R_i) \gtrsim \hat{P}_A|_{\underbrace{AAA\dots}_{h-i+1} \underbrace{BAAA\dots}_{i-1}} (\equiv F_i)$ ), and  $\hat{r}_{\underbrace{AAA\dots}_{h-i+1} \underbrace{BAAA\dots}_{i-1}}^s \ll f_{\underbrace{AAA\dots}_{h-i} \underbrace{BAAA\dots}_i}^s$  for  $m < i \leq h$  (i.e.,  $R_i \ll F_i$ ).  $m$  is an arbitrary integer in the range  $[0, h]$ .

This means that the primer terminus containing a mismatch is more readily transferred and excised rather than extended by the addition of the next matched nucleotide. This makes a significant contribution to the proofreading efficiency. On the other hand, as the mismatch is buried deeper (i.e.,  $i$  gets larger), the transfer-and-excision rate  $\hat{r}_{\underbrace{AAA\dots}_{h-i+1} \underbrace{BAAA\dots}_{i-1}}^s$

decreases and the addition rate  $f_{\underbrace{AAA\dots B}_{h-i} \underbrace{AAA\dots}_{i}}$  increases and far exceeds the transfer-and-excision rate when  $i > m$ . Hence, only those kinetic parameters of  $0 < i \leq m$  contribute significantly to the proofreading efficiency. More details about this condition can be found in Appendix B.

(d)  $f_{X_h\dots X_1 B}^s \approx 0$  (where  $X_h X_{s-1} \dots X_1 \neq \underbrace{AAA\dots}_h$ ), which leads to  $P_{B|X_{h+1}X_h\dots X_1} = 0$ .

This means that the chance of adding one more mismatch within the length of  $h$  is negligible.

With these bio-relevant conditions, a very simple and intuitive expression of the replication fidelity can be obtained:

$$\begin{aligned} \phi &= \phi_s \phi_e, \quad \phi_s = f_{\underbrace{AAA\dots}_{h+1}}^s / f_{\underbrace{AAA\dots}_h}^s B, \\ \phi_e &= \left(1 + \frac{R_1}{F_1}\right) \left(1 + \frac{R_2}{F_2}\right) (\dots) \left(1 + \frac{R_h}{F_h}\right). \end{aligned} \quad (14)$$

$\phi_s, \phi_e$  denotes the contribution of the polymerase pathway and the proofreading pathway to the overall fidelity, respectively (details can be found in Appendix B).

Particularly, for the first-order model, we have,

$$\phi_s = \frac{f_{AA}^s}{f_{AB}^s}, \quad \phi_e = 1 + \frac{\hat{r}_{AB}^s}{f_{BA}^s}. \quad (15)$$

For the second-order model, we have,

$$\phi_s = \frac{f_{AAA}^s}{f_{AAB}^s}, \quad \phi_e = \left(1 + \frac{\hat{r}_{AAB}^s}{f_{ABA}^s}\right) \left(1 + \frac{\hat{r}_{ABA}^s}{f_{BAA}^s}\right). \quad (16)$$

Here  $\hat{r}_{X_3 X_2 X_1}^s = k_{X_3 X_2 X_1}^{se} r_{X_3 X_2 X_1}^e / (r_{X_3 X_2 X_1}^e + k_{X_3 X_2 X_1}^{es})$ , similarly defined as  $\hat{r}_{X_2 X_1}^s$  for the first-order model. If all the parameters are taken as first order, the term  $\hat{r}_{ABA}^s / f_{BAA}^s$  becomes negligible (according to the condition (b)  $f_{AA}^s \gg \hat{r}_{BA}^s$ ), and Eq. (16) is indeed reduced to Eq. (15).

For Model II, following similar procedure, one can derive the same expression of  $\phi$  as Eq. (14) under the same conditions (details see Appendix C). Furthermore, by numerically solving the steady-state kinetic equations (e.g., Eqs. (5)), one can also show that Model I and II give almost the same overall reaction velocity ( $J_{tot} = J_A + J_B$ ) under the bio-relevant conditions (data not shown). This is conceivable, since the overall velocity is dominated by the addition of  $A$  ( $f_{\underbrace{AAA\dots}_{h+1}}^s$  is far larger than any other kinetic parameters) and introduction of proofreading pathway only slightly changes the overall velocity. Therefore, the two models

behave almost the same in steady state under the bio-relevant conditions (they do differ under other conditions, which beyond the topic of the present paper). This means that the details how the excised terminus returns to the polymerase site may be unimportant for real DNAPs to obtain high proofreading efficiency while maintain high polymerization velocity.

#### IV. CASE STUDIES

In the above expressions of  $\phi$ , only a few key parameters appear, which enables us to evaluate the fidelity of some real DNAPs even if other unimportant kinetic parameters are unknown or not precisely measured. Here we give two case studies.

##### A. First-order proofreading

Employing the pre-steady-state kinetic analysis method, K.A.Johnson et al. analyzed the polymerization process and the excision process of T7 DNA polymerase<sup>12,25,30</sup>. The kinetic parameters they obtained are listed in TABLE I, and can be understood as first order parameters. Since they satisfy the bio-relevant conditions, Eq. (15) can be applied here.

For  $\phi_s$ , K.A.Johnson et al. used an expression exactly the same as ours ( $= f_{AA}^s/f_{AB}^s \simeq 10^5$ ). However, for  $\phi_e$ , they calculated as

$$\phi_e = 1 + \frac{k_{AB}^{se}}{f_{BA}^s} \simeq 193. \quad (17)$$

Compared to Eq. (15), it's obvious that they ignored the bidirectional transfer of the primer terminus between the polymerase site and exonuclease site. By our theory, it can be modified as

$$\phi_e \simeq 1 + \frac{\hat{r}_{AB}^s}{f_{BA}^s} = 1 + \frac{k_{AB}^{se}\sigma}{f_{BA}^s} \simeq 107. \quad (18)$$

Here  $\sigma = r^e/(r^e + k_{AB}^{es}) = 0.56$ , not far from its upper limit at which K.A.Johnson et al.'s expression is recovered. Notice that  $\sigma$  could play a negative role if  $\sigma \ll 1$  (i.e.,  $r^e \ll k_{AB}^{es}$ ),  $\sigma = 0.56$  implies that the excision process is highly efficiently employed by T7 DNAP for the proofreading purpose.

To further validate the approximate expression Eq. (15) for T7 DNAP, we compared the approximate result  $\phi_{appr}$  to the exact numerical solution of Eq. (5)  $\phi$  in a large range of the



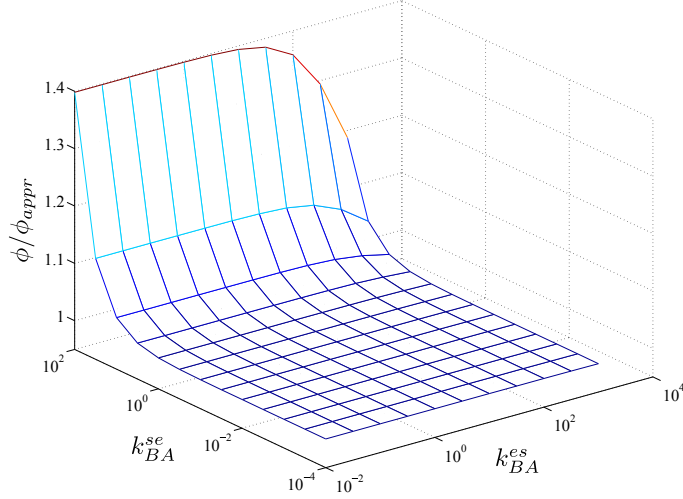


FIG. 8. The fidelity  $\phi$  of T7 DNAP, which is calculated by the exact numerical solution, divided by the approximate expression  $\phi_{appr}$ . It shows that  $\phi$  can be well approximated by  $\phi_{appr}$  in large range of the two undetermined parameters  $k_{BA}^{se}$  and  $k_{BA}^{es}$ . Kinetic parameters are taken from TABLE I (in unit  $s^{-1}$ ):  $f_{AA}^s = 250$ ,  $f_{AB}^s = 0.002$ ,  $f_{BA}^s = 0.012$ ,  $k_{AA}^{se} = 0.2$ ,  $k_{AB}^{se} = 2.3$ ,  $k_{AA}^{es} = 714$ ,  $k_{AB}^{es} = 714$ ,  $r^e = 896$ . All other parameters involving  $BB$  are set as zero.

two undetermined parameters  $k_{BA}^{se}$  and  $k_{BA}^{es}$ . As shown in FIG. 8, both methods give very close results in large range of  $k_{BA}^{se}$  and  $k_{BA}^{es}$ .

## B. Second-order proofreading

For human mitochondrial DNAP pol  $\gamma$ , K.A.Johnson et al. measured some kinetic parameters<sup>26</sup> (TABLE I) which displays the second-order terminal effect. Although some involved parameters have not been determined directly, some of their combinations were measured. For instance,  $\hat{r}_{AAB}^s (= k_{AAB}^{se} r^e / (r^e + k_{AAB}^{es})) = 0.4 s^{-1}$  (i.e.,  $k_{exo} = 0.4 s^{-1}$  in Scheme 1 of Ref. 26),  $\hat{r}_{ABA}^s (= k_{ABA}^{se} r^e / (r^e + k_{ABA}^{es})) = 3 s^{-1}$  (i.e.,  $k_{exo} = 3 s^{-1}$  in Scheme 1 of Ref. 26).

Assuming that above-mentioned bio-relevant conditions are satisfied and using the available kinetic parameters, one can make a rough estimate of the overall fidelity  $\phi = \phi_s \phi_e$  as

follows:

$$\phi_s \simeq \frac{f_{AAA}^s}{f_{AAB}^s} = \frac{3900 - 5700}{0.023 - 1.6} \simeq 10^4 - 10^5, \quad (19a)$$

$$\phi_e \simeq (1 + \frac{\hat{r}_{AAB}^s}{f_{ABA}^s})(1 + \frac{\hat{r}_{ABA}^s}{f_{BAA}^s}) = (1 + \frac{0.4}{0.1})(1 + \frac{3}{2.7}) \simeq 10. \quad (19b)$$

In their article<sup>26</sup>, K.A.Johnson et al. divided  $\phi_e$  intuitively into two multiplying parts. One is due to the correction of the terminal mismatch, and the other is due to the correction of the buried mismatch. In our terminology, they actually considered  $(\hat{r}_{AAB}^s/f_{ABA}^s)$  and  $(\hat{r}_{ABA}^s/f_{BAA}^s)$ , respectively. So their expression of  $\phi_e$  is almost the same as ours Eq. (16).

One may notice that the second-order proofreading contribution  $(\hat{r}_{ABA}^s/f_{BAA}^s)$ , seems insignificant. Fortunately, it can be enhanced, when free dNTP matching the terminal or penultimate base on the template are presented in the solution. Actually, these dNTPs were observed to apparently accelerate the excision of the penultimate mismatch (see Scheme 2 of Ref. 26). This can be understood by the above expression of  $\phi_e$ . Since the duplex terminus is unstable due to the buried mismatch, the free dNTP has the chance to bind transiently to the template at the polymerase site, which may accelerate the transfer of the primer terminus from the polymerase site to the exonuclease site, or hinder the back transfer. This will increase  $k_{ABA}^{se}$  or decrease  $k_{ABA}^{es}$  (in either case, to increase  $\hat{r}_{ABA}^s$ ), and thus enhance  $\phi_e$ . In fact,  $\hat{r}_{ABA}^s$  was found to increase from  $3 \text{ s}^{-1}$  (in the absence of matching dNTP in the solution) to up to  $21 \sim 39 \text{ s}^{-1}$  (in the presence of matching dNTP), which leads to an order of magnitudes increase of  $\phi_e$ .

## V. DISCUSSION AND CONCLUSION

In this work, we propose a general kinetic framework to analyze the fidelity problem of DNAP which owns both a polymerase site for primer growth and an exonuclease site for proofreading. So far as we know, it's the first time that the two sub-processes, as well as the higher-order terminal effects, can be rigorously studied in a unified way (either for Model I or for Model II). Closed equations were derived which fully describe the steady-state replication process. By these equations, the replication fidelity  $\phi$ , as well as other quantities such as the total flux  $J$  (the overall reaction velocity), can be calculated. In particular, using the infinite-state Markov chain method which is numerically equivalent to our steady-state equations, we derived analytical expressions of  $\phi$  for both Model I and Model II under bio-relevant

conditions. We found that Model I and Model II behave almost the same in every aspect (e.g., the fidelity, the overall reaction velocity, etc.) under those conditions. This implies that the proofreading efficiency of DNAP may not depend on the details of how the excised primer terminus returns from the exonuclease site to the polymerase site. Furthermore, the highly simplified expressions of  $\phi$  show that the replication fidelity is only determined by very few kinetic parameters, which indicates that the polymerization-proofreading mechanism is insensitive to details of the reaction schemes.

The expression of  $\phi$  of  $h^{th}$ -order model (Eq. (14)) offers intuitive and important insights to understand the higher-order terminal effects. We noticed that the polymerase site can add  $A$  to the primer terminus with a much larger rate than adding  $B$ , which contributes significantly to the overall fidelity. In this pathway, however, the  $h^{th}$ -order terminal effects are not reflected explicitly in  $\phi_s$ . In fact, the higher-order effects work in the proofreading pathway.

To simply put, when the primer terminus contains one  $B$  at whatever position, it can be extended one  $A$  by the polymerase site, or be transferred and excised by the exonuclease site. Once the former is much larger than the latter (see condition (c), for  $0 < i \leq m$ ), it can substantially contribute to  $\phi$  as a ratio between these two rates. In principle, for each possible position (the terminal, the penultimate, etc.) of  $B$ , there is a corresponding ratio contributing to  $\phi_e$ . However, it seems only a few leading ratios contribute significantly to  $\phi_e$ . As pointed out in Ref. 51, the higher-order effects may originate mainly from base-stacking interaction in the DNA duplex. The presence of terminal or penultimate mismatch may significantly disrupt the base stacking of the duplex terminus, and thus increases the transfer-and-excision rate and decreases the addition rate, which enhances the proofreading contribution to the overall fidelity. On the other hand, deeper mismatches may put less impact on both rates and thus on the proofreading efficiency (see condition (c), for  $m < i \leq h$ ). For instance, in the case of human mitochondrial DNAP pol  $\gamma$ , it has been observed  $f_{BAA}^s \gg f_{ABA}^s$  (TABLE I), and thus the contribution of the buried mismatch (in the absence of matching dNTP in the solution) to  $\phi_e$ ,  $(\hat{r}_{ABA}^s/f_{BAA}^s)$ , is smaller than that of the terminal mismatch  $(\hat{r}_{AAB}^s/f_{ABA}^s)$ . This raises the question that whether the third-order even higher-order effects can be observed for any real DNAPs. For the third-order model,  $\phi_e \simeq (1 + r_{AAAB}^s/f_{AABA}^s)(1 + r_{AABA}^s/f_{ABAA}^s)(1 + r_{ABAA}^s/f_{BAAA}^s)$ . If  $f_{BAAA}^s$  approaches to  $f_{AAAA}^s$  which is much larger than any other kinetic parameter, then the term correspond-

ing to the third-order effect  $r_{ABAA}^s/f_{BAAA}^s$  is negligible, meaning that this has no practical contributions to fidelity. Whether such higher-order effects exist for other DNAPs is worthy investigation in the future.

It should also be pointed out that we have not discuss the sequence effect on the fidelity in this paper. As shown by experiments such as Ref. 26, the 16 possible base pairs may have different incorporation rates or excision rates, which of course have more or less impact on the overall fidelity. Our model and indeed all the existing models, are actually based on the presumption that the rates of the 4 matches are of similar order of magnitude, and the rates of the 12 mismatches are also of similar order of magnitude. This coarse-grained description of DNA replication process is appropriate for the purpose to estimate the overall replication fidelity, but cannot account for much subtle effects such as sequence-dependent replication errors which are biologically very important. To develop a new kinetic theory to take account of the sequence effects would be very challenging, since the symbolic sequence of the template is inevitably involved, which leads to many difficulties for theoretical studies (e.g., it's hard to rigorously define the steady state). Actually, there have been some numerical simulations in that respect (e.g., see Ref. 52), but rigorous modeling of the simulated processes is still lacking. The approach presented in this paper may serve as a start for further development of such a modeling framework.

## ACKNOWLEDGMENTS

The authors thank the financial support by the National Basic Research Program of China (973 program, No.2013CB932804) and National Natural Science Foundation of China (No.11105218, No.91027046, No.11574329 and No.11322543).

## Appendix A: The iterative equation

In the main text, we use  $P_{euX_2X_1}$  to denote the possibility that the newly incorporated  $X_1$  ever being excised. It can be calculated by counting all the possible routes that lead to the finally excision.

(a) The possibility that the terminal  $X_1$  is excised without subsequent dNTP addition can be calculated as  $P_{euX_2X_1}^{00} = P_{X_2X_1}^{se} T_{X_2X_1} P_{u|X_2X_1} = \hat{P}_{u|X_2X_1}$ . Here  $T_{X_2X_1} = 1 + P_{X_2X_1}^{es} P_{X_2X_1}^{se} +$

$(P_{X_2X_1}^{es} P_{X_2X_1}^{se})^2 + \dots = 1/(1 - P_{X_2X_1}^{se} P_{X_2X_1}^{es})$ .  $P_{X_2X_1}^{se} T_{X_2X_1}$  is the sum of the possibility of all the routes that the primer terminus is initially in the polymerase site and then transferred back-and-forth and eventually located at the exonuclease site.

(b) It's also possible that the primer terminal  $X_1$  is buried by the next dNTP addition and eventually be excised. For example,  $X_1$  is buried by the subsequent addition of  $A$  (with a possibility  $T_{X_2X_1} P_{A|X_2X_1}$ ), and this newly added  $A$  is excised (with a possibility  $P_{euX_1A}$ ), and finally  $X_1$  itself is excised (with a possibility  $T_{X_2X_1} P_{u|X_2X_1}$ ). According to this logic, the possibility of the route that  $A$  is incorporated and excised  $i$  times and  $B$  is incorporated and excised  $j$  times before the final excision of  $X_1$ , can be calculated as  $P_{euX_2X_1}^{ij} = C_{i+j}^i (T_{X_2X_1} P_{A|X_2X_1} P_{euX_1A})^i (P_{X_2X_1}^{es})^{i+j-1} (T_{X_2X_1} P_{B|X_2X_1} P_{euX_1B})^j (T_{X_2X_1} P_{u|X_2X_1})$  ( $i + j \geq 1$ ).

Accordingly, we have

$$\begin{aligned}
P_{euX_2X_1} &= \sum_{i,j \geq 0} P_{euX_2X_1}^{ij} \\
&= \hat{P}_{u|X_2X_1} + \frac{T_{X_2X_1} P_{u|X_2X_1}}{P_{X_2X_1}^{es}} \sum_{i+j \geq 1} C_{i+j}^i (\hat{P}_{A|X_2X_1} P_{euX_1A} P_{X_2X_1}^{es})^i (\hat{P}_{B|X_2X_1} P_{euX_1B} P_{X_2X_1}^{es})^j \\
&= \hat{P}_{u|X_2X_1} + \frac{\hat{P}_{u|X_2X_1}}{P_{X_2X_1}^{se} P_{X_2X_1}^{es}} \sum_{n \geq 1} (\hat{P}_{A|X_2X_1} P_{euX_1A} P_{X_2X_1}^{es} + \hat{P}_{B|X_2X_1} P_{euX_1B} P_{X_2X_1}^{es})^n \\
&= \hat{P}_{u|X_2X_1} + \frac{\hat{P}_{u|X_2X_1}}{P_{X_2X_1}^{se} P_{X_2X_1}^{es}} \left( \frac{1}{1 - (\hat{P}_{A|X_2X_1} P_{euX_1A} P_{X_2X_1}^{es} + \hat{P}_{B|X_2X_1} P_{euX_1B} P_{X_2X_1}^{es})} - 1 \right) \\
&= \frac{\hat{P}_{u|X_2X_1}}{P_{X_2X_1}^{se} P_{X_2X_1}^{es}} \left( \frac{1}{1 - (\hat{P}_{A|X_2X_1} P_{euX_1A} + \hat{P}_{B|X_2X_1} P_{euX_1B}) P_{X_2X_1}^{es}} - \frac{1}{T_{X_2X_1}} \right). \tag{A1}
\end{aligned}$$

For higher-order terminal models, one can also obtain recursion equations of the same form.

## Appendix B: The approximation of $\phi$ under bio-relevant conditions

We use the second-order model to demonstrate the approximation.

Under bio-relevant conditions, the fidelity expression can be approximated as,

$$\begin{aligned}
\phi &= \frac{Q_{AAA} + Q_{ABA} + Q_{BAA} + Q_{BBA}}{Q_{AAB} + Q_{ABB} + Q_{BAB} + Q_{BBB}} \\
&\simeq \frac{Q_{AAA} + Q_{ABA} + Q_{BAA}}{Q_{AAB}} \\
&\simeq \frac{Q_{AAA}}{Q_{AAB}} + 2 \simeq \frac{\hat{P}_{AAA} P_{nuAAA}}{\hat{P}_{AAB} P_{nuAAB}}. \tag{B1}
\end{aligned}$$

In the first step, we have  $Q_{ABB} = Q_{BAB} = Q_{BBA} = Q_{BBB} = 0$  because of condition (d). In the second step, we have  $Q_{AAA} \gg Q_{AAB}$  because of the conditions (a) and (b), and  $Q_{ABA} = Q_{BAA} = Q_{AAB}$  due to constraint Eq. (12) (i.e.,  $Q_{XY} = Q_{AXY} + Q_{BXY} = Q_{XYA} + Q_{XYB}$ ).

The fidelity expression can then be separated into two parts,  $\phi_s = \hat{P}_{AAA}/\hat{P}_{AAB}$  and  $\phi_e = P_{nuAAA}/P_{nuAAB}$ . The first part is the contribution of polymerase site, which can be easily calculated as  $\phi_s = \hat{P}_{AAA}/\hat{P}_{AAB} = f_{AAA}^s/f_{AAB}^s$ . The second part  $\phi_e$  is the contribution of exonuclease site, which can be calculated as follows.

First,  $P_{nuAAA} = 1 - P_{euAAA} \simeq 1$ , since  $P_{euAAA} \simeq 0$  (this is intuitive according to conditions (a) and (b), and can be verified by numerical calculation). Thus, the fidelity  $\phi_e$  is determined by  $P_{nuAAB} = 1 - P_{euAAB}$ . For  $P_{euAAB}$ , similar to Appendix A, we have

$$\begin{aligned} P_{euAAB} &= \frac{\hat{P}_{u|AAB}}{P_{AAB}^{se} P_{AAB}^{es}} \left( \frac{1}{1 - (\hat{P}_{A|AAB} P_{euABA} + \hat{P}_{B|AAB} P_{euABB}) P_{AAB}^{es}} - \frac{1}{T_{AAB}} \right) \\ &= \frac{\hat{P}_{u|AAB}}{P_{AAB}^{se} P_{AAB}^{es}} \left( \frac{1}{1 - (\mathcal{A}_1^{(2)} + \mathcal{B}_1^{(2)}) P_{AAB}^{es}} - \frac{1}{T_{AAB}} \right) \\ &= \frac{\hat{P}_{u|AAB}}{P_{AAB}^{se} P_{AAB}^{es}} \left( \frac{1}{1 - \mathcal{A}_1^{(2)} P_{AAB}^{es}} - \frac{1}{T_{AAB}} \right), \end{aligned} \quad (B2)$$

where  $\mathcal{A}_1^{(2)} = \hat{P}_{A|AAB} P_{euABA}$ ,  $\mathcal{B}_1^{(2)} = \hat{P}_{B|AAB} P_{euABB}$ . In the second step, we used  $\mathcal{B}_1^{(2)} = 0$  because of condition (d). Now all the quantities in the expression of  $P_{euAAB}$  are known except  $P_{euABA}$  which can be expressed as,

$$\begin{aligned} P_{euABA} &= \frac{\hat{P}_{u|ABA}}{P_{ABA}^{se} P_{ABA}^{es}} \left( \frac{1}{1 - (\hat{P}_{A|ABA} P_{euBAA} + \hat{P}_{B|ABA} P_{euBAB}) P_{ABA}^{es}} - \frac{1}{T_{ABA}} \right) \\ &= \frac{\hat{P}_{u|ABA}}{P_{ABA}^{se} P_{ABA}^{es}} \left( \frac{1}{1 - (\mathcal{A}_2^{(2)} + \mathcal{B}_2^{(2)}) P_{ABA}^{es}} - \frac{1}{T_{ABA}} \right) \\ &= \frac{\hat{P}_{u|ABA}}{P_{ABA}^{se} P_{ABA}^{es}} \left( \frac{1}{1 - \mathcal{A}_2^{(2)} P_{ABA}^{es}} - \frac{1}{T_{ABA}} \right), \end{aligned} \quad (B3)$$

where  $\mathcal{A}_2^{(2)} = \hat{P}_{A|ABA} P_{euBAA}$ ,  $\mathcal{B}_2^{(2)} = \hat{P}_{B|ABA} P_{euBAB}$ . In the second step, we used  $\mathcal{B}_2^{(2)} = 0$  because of condition (d). As for  $\mathcal{A}_2^{(2)}$ , it's actually negligible. To make it clear, we resort to

the expression of  $P_{euBAA}$ :

$$\begin{aligned}
P_{euBAA} &= \frac{\hat{P}_{u|BAA}}{P_{BAA}^{se} P_{BAA}^{es}} \left( \frac{1}{1 - (\hat{P}_{A|BAA} P_{euAAA} + \hat{P}_{B|BAA} P_{euAAB}) P_{BAA}^{es}} - \frac{1}{T_{BAA}} \right) \\
&= \frac{\hat{P}_{u|BAA}}{P_{BAA}^{se} P_{BAA}^{es}} \left( \frac{1}{1 - (\mathcal{A}_3^{(2)} + \mathcal{B}_3^{(2)}) P_{BAA}^{es}} - \frac{1}{T_{BAA}} \right) \\
&= \hat{P}_{u|BAA} \simeq 0,
\end{aligned} \tag{B4}$$

where  $\mathcal{A}_3^{(2)} = \hat{P}_{A|BAA} P_{euAAA}$ ,  $\mathcal{B}_3^{(2)} = \hat{P}_{B|BAA} P_{euAAB}$ .  $\mathcal{A}_3^{(2)} \simeq 0$  since  $P_{euAAA} = 0$ .  $\mathcal{B}_3^{(2)} \simeq 0$  since  $\hat{P}_{B|BAA} \simeq 0$  because of condition (a). Finally, we obtain  $P_{euBAA} \simeq \hat{P}_{u|BAA} \simeq 0$  because of condition (a) and (b).

Now we have  $P_{euABA} \simeq \hat{P}_{u|ABA}$ ,  $\mathcal{A}_1^{(2)} = \hat{P}_{A|AAB} \hat{P}_{u|ABA}$ , and

$$P_{euAAB} \simeq \frac{\hat{P}_{u|AAB}}{P_{AAB}^{se} P_{AAB}^{es}} \left( \frac{1}{1 - \hat{P}_{A|AAB} \hat{P}_{u|ABA} P_{AAB}^{es}} - \frac{1}{T_{AAB}} \right). \tag{B5}$$

This expression is of the following general form:

$$p_1 = \frac{\alpha}{\theta\gamma} \left( \frac{1}{1 - (1 - \alpha)\beta\gamma} - (1 - \theta\gamma) \right), \tag{B6}$$

where  $\alpha = \theta(1 - \gamma)/(1 - \theta\gamma)$  and  $0 < \alpha, \beta, \gamma, \theta < 1$ . It can be approximated by the following simpler expression:

$$p_2 = \alpha + (1 - \alpha)\beta. \tag{B7}$$

$p_1 \simeq p_2$  holds for  $\alpha \gtrsim 0.5$  (here it means  $\hat{P}_{u|AAB} \gtrsim 0.5$ , i.e.,  $R_i \gtrsim F_i$  for  $0 < i \leq 1$ , see condition (c)), which can be verified numerically as shown by FIG. 9. Thus, we can write  $P_{euAAB}$  as,

$$P_{euAAB} \simeq \hat{P}_{u|AAB} + \hat{P}_{A|AAB} \hat{P}_{u|ABA}. \tag{B8}$$

The fidelity  $\phi_e \simeq 1/(1 - P_{euAAB})$  can then be calculated. Finally we obtain an intuitive approximate expression of the overall replication fidelity:

$$\phi \simeq \frac{\hat{P}_{AAA}}{\hat{P}_{AAB}(1 - P_{euAAB})} = \frac{\hat{P}_{AAA}}{\hat{P}_{AAB} \hat{P}_{A|AAB} \hat{P}_{A|ABA}} = \frac{f_{AAA}^s}{f_{AAB}^s} \left( 1 + \frac{\hat{r}_{AAB}^s}{f_{ABA}^s} \right) \left( 1 + \frac{\hat{r}_{ABA}^s}{f_{BAA}^s} \right). \tag{B9}$$

It should be noted that under conditions that  $\alpha \ll 1$  and  $\beta \ll 1$  (i.e.,  $R_i \ll F_i$  for  $0 < i \leq 2$ , see condition (c)), the expression Eq. (B9) is still valid, since  $p_1 \simeq 0$  and thus  $\phi_e \simeq 1$ .

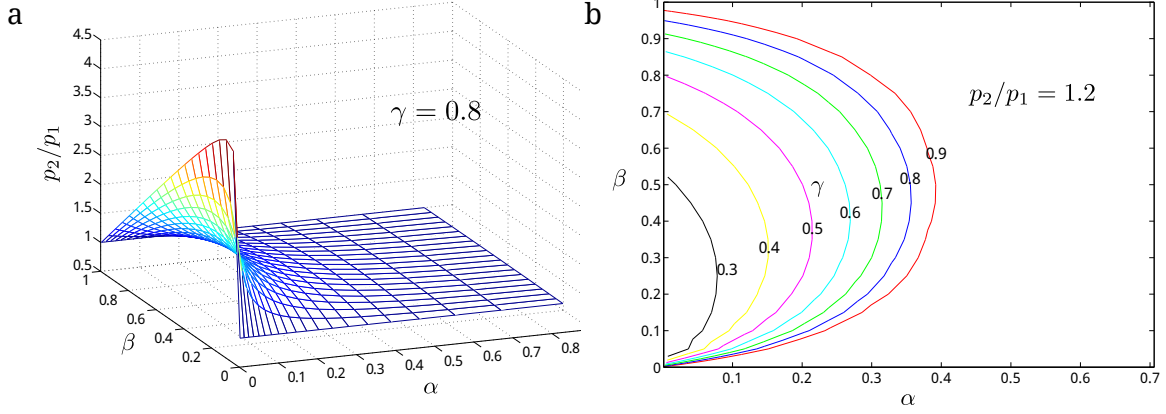


FIG. 9. (a) Comparison between  $p_2$  and  $p_1$  in the parameter space  $(\alpha, \beta)$ .  $\gamma$  is set to 0.8. (b) The isolines of  $p_2/p_1 = 1.2$ , for  $\gamma$  varying from 0.3 to 0.9. Right to all isolines is the range of  $\alpha$  and  $\beta$  in which  $p_2/p_1$  almost equals to 1.

Extending this logic to higher-order models is straightforward. For  $h^{th}$ -order model, we similarly have

$$\begin{aligned} \phi &\equiv \frac{\sum_{X_i=A,B} \hat{P}_{X_{h+1}X_h \dots X_2 A} P_{nu X_{h+1}X_h \dots X_2 A}}{\sum_{X_i=A,B} \hat{P}_{X_{h+1}X_h \dots X_2 B} P_{nu X_{h+1}X_h \dots X_2 B}} \simeq \frac{\hat{P}_{\underbrace{AAA \dots}_{h+1}} P_{nu \underbrace{AAA \dots}_{h+1}}}{\hat{P}_{\underbrace{AAA \dots}_h B} P_{nu \underbrace{AAA \dots}_h B}} + h \\ &\simeq \frac{\hat{P}_{\underbrace{AAA \dots}_{h+1}} P_{nu \underbrace{AAA \dots}_{h+1}}}{\hat{P}_{\underbrace{AAA \dots}_h B} P_{nu \underbrace{AAA \dots}_h B}}, \end{aligned} \quad (B10)$$

and we also have  $P_{eu \underbrace{AAA \dots}_{h+1}} \simeq 0$ , thus  $P_{nu \underbrace{AAA \dots}_{h+1}} \simeq 1$ . To calculate  $P_{nu \underbrace{AAA \dots}_h B} = 1 - P_{eu \underbrace{AAA \dots}_h B}$ , we have to calculate all the following,

$$R_i^* \equiv P_{eu \underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}} \simeq \frac{\hat{P}_{u| \underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}}}{P_{se \underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}} P_{es \underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}}} \left( \frac{1}{1 - (\mathcal{A}_i^{(h)} + \mathcal{B}_i^{(h)}) P_{es \underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}}} - \frac{1}{T_{\underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}}} \right),$$

$1 \leq i \leq h+1$ , where  $\mathcal{A}_i^{(h)} \equiv \hat{P}_{A| \underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}} P_{eu \underbrace{AAA \dots}_{h-i} B \underbrace{AAA \dots}_i}$ ,  $\mathcal{B}_i^{(h)} \equiv \hat{P}_{B| \underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}} P_{eu \underbrace{AAA \dots}_{h-i} B \underbrace{AAA \dots}_i}$

for  $1 \leq i \leq h$ , and  $\mathcal{A}_{h+1}^{(h)} \equiv \hat{P}_{A|B \underbrace{AAA \dots}_h} P_{eu \underbrace{AAA \dots}_{h+1}}$ ,  $\mathcal{B}_{h+1}^{(h)} \equiv \hat{P}_{B|B \underbrace{AAA \dots}_h} P_{eu \underbrace{AAA \dots}_{h+1}}$ . We have

$\mathcal{B}_i^{(h)} \simeq 0$  ( $1 \leq i \leq h$ ) because of condition (d),  $\mathcal{B}_{h+1}^{(h)} \simeq 0$  because of condition (a), and  $\mathcal{A}_{h+1}^{(h)} \simeq 0$  since  $P_{eu \underbrace{AAA \dots}_{h+1}} \simeq 0$ . So we obtain  $R_{h+1}^* \simeq \hat{P}_{u|B \underbrace{AAA \dots}_h} \simeq 0$  because of condition (a) and (b).



As defined in the main text,  $F_i \equiv \hat{P}_A | \underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}$ , and  $R_i \equiv \hat{P}_u | \underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}$ . For  $1 \leq i \leq h$ , they can be written as (due to condition (d)):

$$F_i = \frac{f_{\underbrace{AAA \dots}_{h-i} B \underbrace{AAA \dots}_i}^s}{f_{\underbrace{AAA \dots}_{h-i} B \underbrace{AAA \dots}_i}^s + \hat{r}_{\underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}}^s}, \quad (\text{B11a})$$

$$R_i = \frac{\hat{r}_{\underbrace{AAA \dots}_{h-i+1} B \underbrace{AAA \dots}_{i-1}}^s}{f_{\underbrace{AAA \dots}_{h-i} B \underbrace{AAA \dots}_i}^s + \hat{r}_{\underbrace{AAA \dots}_{s-i+1} B \underbrace{AAA \dots}_{i-1}}^s}, \quad (\text{B11b})$$

and obviously  $F_i + R_i = 1$ .

To calculate other  $R_i^*$ , we first notice that  $\mathcal{A}_h^{(h)} \simeq 0$  and thus  $R_h^* \simeq R_h$ , since  $R_{h+1}^* \simeq 0$ . For  $R_i^*$  ( $1 \leq i \leq h-1$ ), we can express it in a form similar to Eq. (B6) and thus it can be rewritten as Eq. (B7) under conditions (c')  $R_i \gtrsim F_i$  ( $1 \leq i \leq h-1$ ):

$$R_i^* = R_i + F_i R_{i+1}^* (1 \leq i \leq h). \quad (\text{B12})$$

Hence, it can be easily proved that

$$\phi_e \simeq \frac{1}{1 - R_1^*} \simeq \frac{1}{F_1} \frac{1}{F_2} \dots \frac{1}{F_h} = (1 + \frac{R_1}{F_1})(1 + \frac{R_2}{F_2})(\dots)(1 + \frac{R_h}{F_h}). \quad (\text{B13})$$

Thus, for  $h^{th}$ -order model, we have,

$$\phi = \phi_s \phi_e \simeq \frac{f_{\underbrace{AAA \dots}_h A}^s}{f_{\underbrace{AAA \dots}_h B}^s} (1 + \frac{R_1}{F_1})(1 + \frac{R_2}{F_2})(\dots)(1 + \frac{R_h}{F_h}). \quad (\text{B14})$$

Similar to the second-order model, it can be seen that this expression is still valid under the condition (c) which is less restrictive and more practical than condition (c').

## Appendix C: Approximation of $\phi$ for Model II

The minimal second-order scheme of Model II is shown as FIG. 10. The effective excision rate  $\hat{r}_{X_3 X_2 X_1}$  is the same as that in Model I, which is  $\hat{r}_{X_3 X_2 X_1}^s = k_{X_3 X_2 X_1}^{se} r_{X_3 X_2 X_1}^e / (r_{X_3 X_2 X_1}^e + k_{X_3 X_2 X_1}^{es})$ . As shown in the supplement of our previous paper<sup>29</sup>, infinite-state Markov chain method can also apply to Model II, and the iterative expression for  $P_{eu X_3 X_2 X_1}$  is,

$$P_{eu X_3 X_2 X_1} = \frac{\hat{P}_{u|X_3 X_2 X_1}}{1 - (\hat{P}_{A|X_3 X_2 X_1} P_{eu X_2 X_1 A} + \hat{P}_{B|X_3 X_2 X_1} P_{eu X_2 X_1 B})}. \quad (\text{C1})$$

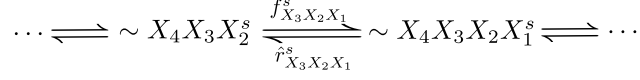


FIG. 10. The minimal second-order reaction scheme for Model II

We also define  $\hat{P}_{u|X_3 X_2 X_1} = \hat{r}_{X_3 X_2 X_1}^s / (\hat{r}_{X_3 X_2 X_1}^s + f_{X_2 X_1 A}^s + f_{X_2 X_1 B}^s)$ , and  $\hat{P}_{X|X_3 X_2 X_1} = f_{X_2 X_1 X}^s / (\hat{r}_{X_3 X_2 X_1}^s + f_{X_2 X_1 X}^s + f_{X_2 X_1 \bar{X}}^s)$ .

Under bio-relevant conditions, the fidelity expression can be approximated as,

$$\begin{aligned} \phi &= \frac{Q_{AAA} + Q_{ABA} + Q_{BAA} + Q_{BBA}}{Q_{AAB} + Q_{ABB} + Q_{BAB} + Q_{BBB}} \\ &\simeq \frac{\hat{P}_{AAA} P_{nuAAA}}{\hat{P}_{AAB} P_{nuAAB}} \\ &= \phi_s \phi_e, \end{aligned} \tag{C2}$$

where  $\phi_s = \hat{P}_{AAA} / \hat{P}_{AAB} = f_{AAA}^s / f_{AAB}^s$ ,  $\phi_e = P_{nuAAA} / P_{nuAAB} \simeq 1 / P_{nuAAB}$ .

Therefore, the fidelity  $\phi_e$  is determined by  $P_{nuAAB} = 1 - P_{euAAB}$ . For  $P_{euAAB}$ ,

$$\begin{aligned} P_{euAAB} &= \frac{\hat{P}_{u|AAB}}{1 - (\hat{P}_{A|AAB} P_{euABA} + \hat{P}_{B|AAB} P_{euABB})} \\ &\simeq \frac{\hat{P}_{u|AAB}}{1 - \hat{P}_{A|AAB} P_{euABA}}. \end{aligned} \tag{C3}$$

$P_{euABA}$  can be calculated as,

$$\begin{aligned} P_{euABA} &= \frac{\hat{P}_{u|ABA}}{1 - (\hat{P}_{A|ABA} P_{euBAA} + \hat{P}_{B|ABA} P_{euBAB})} \\ &\simeq \frac{\hat{P}_{u|ABA}}{1 - \hat{P}_{A|ABA} P_{euBAA}}. \end{aligned} \tag{C4}$$

$P_{euBAA}$  is shown to be negligible:

$$\begin{aligned} P_{euBAA} &= \frac{\hat{P}_{u|BAA}}{1 - (\hat{P}_{A|BAA} P_{euAAA} + \hat{P}_{B|BAA} P_{euAAB})} \\ &\simeq \frac{\hat{P}_{u|BAA}}{1 - \hat{P}_{A|BAA} P_{euAAA}} \\ &\simeq \hat{P}_{u|BAA} \simeq 0. \end{aligned} \tag{C5}$$

So we have  $P_{euABA} \simeq \hat{P}_{u|ABA}$ . Substituting it to Eq. (C3), we obtain

$$P_{euAAB} \simeq \hat{P}_{u|AAB} \left( \frac{1}{1 - \hat{P}_{A|AAB} \hat{P}_{u|ABA}} \right), \tag{C6}$$

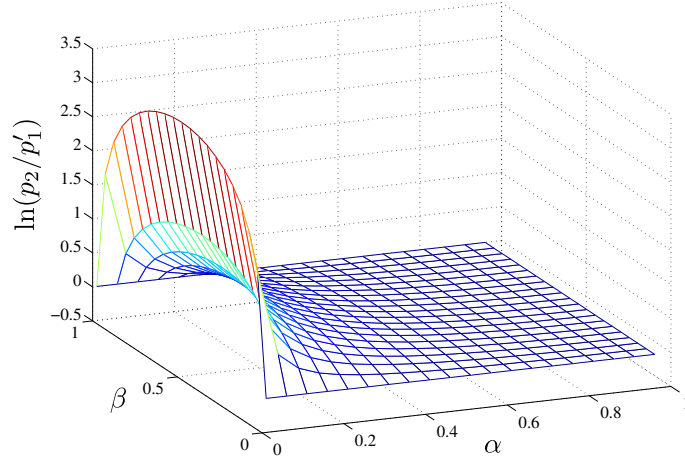


FIG. 11. Comparison between  $p'_1$  and  $p_2$ , in the parameter space  $(\alpha, \beta)$ .

which is of the general form:

$$p'_1 = \alpha \left( \frac{1}{1 - (1 - \alpha)\beta} \right). \quad (\text{C7})$$

Numerical calculations show that Eq. (B7) is also a good approximation of such a function under the condition that  $\alpha \gtrsim 0.5$ , as indicated by FIG. 11. Thus, we obtain the same expression of  $P_{euAAB}$  for both Model I and Mode II:

$$P_{euAAB} \simeq \hat{P}_{u|AAB} + \hat{P}_{A|AAB} \hat{P}_{u|ABA}. \quad (\text{C8})$$

The overall fidelity can be expressed by Eq. (B9). When  $\hat{P}_{u|AAB} \ll 1$  and  $\hat{P}_{u|ABA} \ll 1$ , we have  $P_{euAAB} \simeq \hat{P}_{u|AAB} \simeq 0$ , and the expression is still valid.

Similar to Model I, this result can be directly extended to higher-order models under condition (c') and (c), which gives the same expression as Eq. (B14) in Model I.

## REFERENCES

- <sup>1</sup>J. D. Watson, F. H. Crick, *et al.*, Nature **171**, 737 (1953).
- <sup>2</sup>I. Lehman, M. J. Bessman, E. S. Simms, and A. Kornberg, J. biol. Chem **233**, 163 (1958).
- <sup>3</sup>T. A. Kunkel and K. Bebenek, Annual review of biochemistry **69**, 497 (2000).
- <sup>4</sup>J. J. Hopfield, Proceedings of the National Academy of Sciences **71**, 4135 (1974).
- <sup>5</sup>J. Ninio, Biochimie **57**, 587 (1975).
- <sup>6</sup>D. J. Galas and E. W. Branscomb, Journal of molecular biology **124**, 653 (1978).

- <sup>7</sup>L. K. Clayton, M. F. Goodman, E. W. Branscomb, and D. J. Galas, *Journal of Biological Chemistry* **254**, 1902 (1979).
- <sup>8</sup>K. A. Johnson, *Annual review of biochemistry* **62**, 685 (1993).
- <sup>9</sup>M. F. Goodman, *Proceedings of the National Academy of Sciences* **94**, 10493 (1997).
- <sup>10</sup>M. F. Goodman and D. K. Fygenson, *Genetics* **148**, 1475 (1998).
- <sup>11</sup>A. R. Fersht, *Proceedings of the National Academy of Sciences* **76**, 4946 (1979).
- <sup>12</sup>S. S. Patel, I. Wong, and K. A. Johnson, *Biochemistry* **30**, 511 (1991).
- <sup>13</sup>J. Cline, J. C. Braman, and H. H. Hogrefe, *Nucleic Acids Research* **24**, 3546 (1996).
- <sup>14</sup>G. J. Wuite, S. B. Smith, M. Young, D. Keller, and C. Bustamante, *Nature* **404**, 103 (2000).
- <sup>15</sup>Y.-C. Tsai and K. A. Johnson, *Biochemistry* **45**, 9675 (2006).
- <sup>16</sup>P. Xie, *Journal of theoretical biology* **259**, 434 (2009).
- <sup>17</sup>A. K. Sharma and D. Chowdhury, *Physical Review E* **86**, 011913 (2012).
- <sup>18</sup>K. R. Lieberman, J. M. Dahl, A. H. Mai, A. Cox, M. Akesson, and H. Wang, *Journal of the American Chemical Society* **135**, 9149 (2013).
- <sup>19</sup>K. R. Lieberman, J. M. Dahl, and H. Wang, *Journal of the American Chemical Society* **136**, 7117 (2014).
- <sup>20</sup>D. Ollis, P. Brick, R. Hamlin, N. Xuong, and T. Steitz, *Nature* (1985).
- <sup>21</sup>A. J. Berman, S. Kamtekar, J. L. Goodman, J. M. Lázaro, M. de Vega, L. Blanco, M. Salas, and T. A. Steitz, *The EMBO journal* **26**, 3494 (2007).
- <sup>22</sup>S. Doublié, S. Tabor, A. M. Long, C. C. Richardson, and T. Ellenberger, *nature* **391**, 251 (1998).
- <sup>23</sup>S. Kamtekar, A. J. Berman, J. Wang, J. M. Lázaro, M. de Vega, L. Blanco, M. Salas, and T. A. Steitz, *Molecular cell* **16**, 609 (2004).
- <sup>24</sup>J. Wang, A. A. Sattar, C. Wang, J. Karam, W. Konigsberg, and T. Steitz, *Cell* **89**, 1087 (1997).
- <sup>25</sup>M. J. Donlin, S. S. Patel, and K. A. Johnson, *Biochemistry* **30**, 538 (1991).
- <sup>26</sup>A. A. Johnson and K. A. Johnson, *Journal of Biological Chemistry* **276**, 38097 (2001).
- <sup>27</sup>R. Lamichhane, S. Y. Berezhna, J. P. Gill, E. Van der Schans, and D. P. Millar, *Journal of the American Chemical Society* **135**, 4735 (2013).
- <sup>28</sup>L. S. Beese, V. Derbyshire, and T. A. Steitz, *Science* **260**, 352 (1993).
- <sup>29</sup>Y.-G. Shu, Y.-S. Song, Z.-C. Ou-Yang, and M. Li, *Journal of Physics: Condensed Matter*

- 27**, 235105 (2015).
- <sup>30</sup>I. Wong, S. S. Patel, and K. A. Johnson, *Biochemistry* **30**, 526 (1991).
- <sup>31</sup>A. A. Johnson and K. A. Johnson, *Journal of Biological Chemistry* **276**, 38090 (2001).
- <sup>32</sup>D. T. Gillespie, *The journal of physical chemistry* **81**, 2340 (1977).
- <sup>33</sup>S. J. Johnson and L. S. Beese, *Cell* **116**, 803 (2004).
- <sup>34</sup>K. R. Tindall and T. A. Kunkel, *Biochemistry* **27**, 6008 (1988).
- <sup>35</sup>K. S. Lundberg, D. S. Dan, M. W. W. Adams, J. M. Short, J. A. Sorge, and E. J. Mathur, *Gene* **108**, 1 (1991).
- <sup>36</sup>R. J. Kokoska, K. Bebenek, F. Boudsocq, R. Woodgate, and T. A. Kunkel, *Journal of Biological Chemistry* **277**, 19633 (2002).
- <sup>37</sup>M. S. Boosalis, J. Petruska, and M. Goodman, *Journal of Biological Chemistry* **262**, 14689 (1987).
- <sup>38</sup>M. F. Goodman, S. Creighton, L. B. Bloom, J. Petruska, and T. A. Kunkel, *Critical reviews in biochemistry and molecular biology* **28**, 83 (1993).
- <sup>39</sup>K. Bebenek, C. Joyce, M. P. Fitzgerald, and T. Kunkel, *Journal of Biological Chemistry* **265**, 13878 (1990).
- <sup>40</sup>R. Kuchta, V. Mizrahi, P. Benkovic, K. Johnson, and S. Benkovic, *Biochemistry* **26**, 8410 (1987).
- <sup>41</sup>R. D. Kuchta, P. Benkovic, and S. J. Benkovic, *Biochemistry* **27**, 6716 (1988).
- <sup>42</sup>K. A. Fiala and Z. Suo, *Biochemistry* **43**, 2106 (2004).
- <sup>43</sup>F. Boudsocq, S. Iwai, F. Hanaoka, and R. Woodgate, *Nucleic Acids Research* **29**, 4607 (2001).
- <sup>44</sup>T. Kunkel, A. Hizi, M. Shaharabany, A. Tsygankov, B. Bröker, J. Fargnoli, J. Ledbetter, J. Bolen, M. De Vivo, J. Chen, *et al.*, *J. Biol. Chem* **267** (1992).
- <sup>45</sup>F. Cady and H. Qian, *Physical biology* **6**, 036011 (2009).
- <sup>46</sup>L. B. Bloom, X. Chen, D. K. Fygenson, J. Turner, M. O'Donnell, and M. F. Goodman, *Journal of Biological Chemistry* **272**, 27919 (1997).
- <sup>47</sup>H. Miller and F. W. Perrino, *Biochemistry* **35**, 12919 (1996).
- <sup>48</sup>T. L. Capson, J. A. Peliska, B. F. Kaboord, M. W. Frey, C. Lively, M. Dahlberg, and S. J. Benkovic, *Biochemistry* **31**, 10984 (1992).
- <sup>49</sup>J. A. Esteban, M. Salas, and L. Blanco, *Journal of Biological Chemistry* **268**, 2719 (1993).
- <sup>50</sup>J. A. Esteban, M. S. Soengas, M. Salas, and L. Blanco, *Journal of Biological Chemistry*

- 269**, 31946 (1994).
- <sup>51</sup>J. Petruska, M. F. Goodman, M. S. Boosalis, L. C. Sowers, C. Cheong, and I. Tinoco, Proceedings of the National Academy of Sciences **85**, 6252 (1988).
- <sup>52</sup>D. A. Pierre Gaspard, Journal of Chemical Physics **141** (2014).